Prediction of car sale prices based on mileage of these cars for different brands

1. Introduction

1.1 Aim and Scope of the Work

A specialized dealer of used cars for various brands wants to know the relationship between the mileage of these cars and their sales price. They also want to see the comparison of the slopes and intercepts mileage and price for different brands of cars.

The motivation of this work is to find out the expert analysis and give some recommendations to increase their profitability by setting the right pricing for their car sales business.

1.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process where a number of techniques are used to understand better the dataset being used.

'Understanding the dataset' can refer to a number of things, including but not limited to…

- Extracting important variables and leaving behind useless variables
- Identifying outliers, missing values, or human error
- Understanding the relationship(s), or lack of, between variables
- Ultimately, maximizing your insights into a dataset and minimizing potential error or that may occur later in the process

Components of EDA

There are main components of exploring data:

1. Understanding your variables

2. Cleaning your dataset

3. Analysing relationships between variables

1.2.1 Understanding the Variable

a) shape returns the number of rows by the number of columns for my dataset. My output was (7253, 22), meaning the dataset has 7253 rows and 13 columns.
b) .head() returns the first 5 rows of my dataset. This is useful if you want to see some example values for each variable.
c) .columns returns the name of all of your columns in the dataset.

1.2.2 Cleaning the Dataset

In the data pre-processing process, these missing values were filled using mean values as the describe command shows mostly values range around mean values.

Some columns, such as power and new price, were dropped as these columns were not relevant. After these data pre-processing values, all the missing values were handled.

1.2.2 Analysing the Relationship between variables

a) Selection of input parameters using Pearson correlation coefficient

The selection of the input variables is a crucial step in the forecasting process because of the complexity of the input values. The choice of input variables can be made using a variety of statistical methods. In this study, the Pearson correlation coefficient was used to select the appropriate input parameters for the prediction of the car price. The Pearson correlation coefficient (r) is one of the simplest and quickest approaches for feature selection. It is defined as the ratio among the covariance of the two features and the standard deviation of them as indicated in the following Equation 1:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (1)$$

- After calculating the correlation coefficient between the price and other features, the values with higher correlation coefficients were observed.
- The highest values of correlation coefficient observed between engine (0.60), mileage (-0.28), and year (0.27).

b) Scatter plots

It's pretty hard to beat correlation heat maps when it comes to data visualizations, but scatterplots are arguably one of the most useful visualizations when it comes to data.

A scatterplot is a type of graph that 'plots' the values of two variables along two axes, like age and height. Scatterplots are useful for many reasons: like correlation matrices, it allows you to quickly understand a relationship between two variables, it's useful for identifying outliers, and they are instrumental when polynomial multiple regression models (which we'll get to in the next article). I used **.plot ()** and set the 'kind' of the graph as **the scatter.** I also set the x-axis as 'odometer' and the y-axis as 'price' since we want to see how different levels of mileage affect the price.

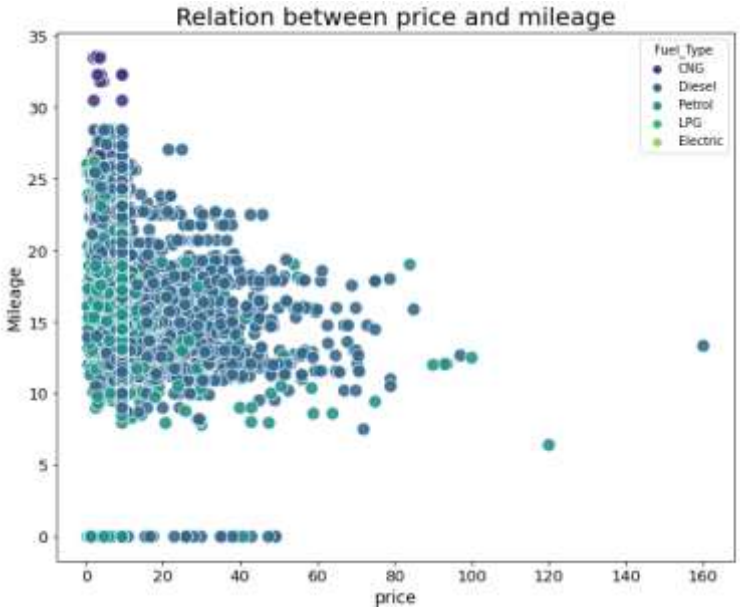

Fig1. Relation between price and mileage of car for different fuel type

2. Regression Analysis

a) Machine learning-based multiple linear regression model

The linear regression method is usually applied to the dataset for the best forecasting results. It is generally used to forecast the linear equation coefficient when it comprises one or more independent variables and is helpful for predicting the dependent variable. This method determines the correlation between several independent variables and a dependent variable. In this study, the MLR model is applied as a suitable technique to deal with the given dataset. The most common equation of the MLR model was used as given in Equation 2:

$$y_n = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \quad (2)$$

Here

y_n denotes the n^{th} dependent variable value
x_n denotes the n^{th} independent variable value
b_0 denotes the intercept of the equation
b_n denotes the n^{th} regression coefficient

b) Development of the model

The data has been divided into two categories, i.e., training and testing data, with 70% of the data utilized for the training and 30% of the data used for testing purposes to ensure the model's accuracy.

c) Model evaluation

The performance of the MLR model can be evaluated using an indicator termed the coefficient of determination (R^2) to determine the model's accuracy. Here the R^2 signifies the forecasting ability. R^2 values lie in the range of 0 to 1, with the higher value of R^2 indicating better forecasting abilities by the MLR method.
The following Equation 3 can be used to evaluate the indicator:

$$R^2 = \frac{(\sum_{i=1}^{N}(y_p^i - \overline{y_p})(y_o^i - \overline{y_o}))^2}{\sum_{i=1}^{N}(y_p^i - \overline{y^i})^2 \sum_{i=1}^{N}(y_o^i - \overline{y_o})} \quad (3)$$

Here,

y_p^i and y_o^i denote the i^{th} predicted and observed values;

$\overline{y_p}$ and $\overline{y_o}$ denote the average of the forecasted and measured values;

and n indicates the number of samples.

3. Discussion

- Initially, simple linear regression was applied between price as the dependent variable and mileage as the independent variable.
- Then, simple linear regression was applied between price as the dependent variable and engine as the independent variable.
  - After measuring the efficiency of the above-applied model, the multiple linear regression model was applied.
  - In the MLR model, price as the dependent variable and mileage, engine, and year as the independent variable were applied.

Final Results

- When a simple linear regression model was applied between the dependent and independent variables, the values of RMSE were 9.38, and the R2 score was 0.07.

- When a multiple linear regression model was applied between dependent price and independent variable as engine, mileage, and year the value RMSE was 6.91, and the R2 score was 0.50.

4. Limitations

1. Regression models cannot work properly if the input data has errors (that is, poor-quality data). If the data pre-processing is not performed well to remove missing values, redundant data or outliers, or imbalanced data distribution, the validity of the regression model suffers.

2. Regression models are susceptible to collinear problems (that is, there exists a strong linear correlation between the independent variables). If the independent

variables are strongly correlated, then they will eat into each other's predictive power, and the regression coefficients will lose their ruggedness.

3. As the number of variables increases, the reliability of the regression models decreases. The regression models work better if you have a small number of variables.

4. Regression models do not automatically take care of nonlinearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit.

5. Regression models work with datasets containing numeric values and not with categorical variables. There are ways to deal with categorical variables though, by creating multiple new variables with a yes/no value.

5. Conclusion

- When the simple linear regression model was applied the efficiency of model as RMSE values was 9.38 which was more when multiple linear regression model was applied the efficiency of model as RMSE values was 6.91.

- When multiple linear regression model was applied the above model give highest efficiency as it gives lowest RMSE value with good R2 score.

- The efficiency of the above model can also be improved using data handling and cleansing process.

- The efficiency of the above model can also be improved by incorporating more dataset.